GroupLiNGAM: Linear non-Gaussian acyclic models for sets of variables

Yoshinobu Kawahara,[1] Kenneth Bollen,[2] Shohei Shimizu[1] and Takashi Washio[1]

Abstract: Finding the structure of a graphical model has been received much attention in many fields. Recently, it is reported that the non-Gaussianity of data enables us to identify the structure of a directed acyclic graph without any prior knowledge on the structure. In this paper, we propose a novel non-Gaussianity based algorithm for more general type of models; chain graphs. The algorithm finds an ordering of the disjoint subsets of variables by iteratively evaluating the independence between the variable subset and the residuals when the remaining variables are regressed on those. However, its computational cost grows exponentially according to the number of variables. Therefore, we further discuss an efficient approximate approach for applying the algorithm to large sized graphs. We illustrate the algorithm with artificial and real-world datasets.

1. Introduction

In order to discover or understand data generating mechanisms, a graphical model has been used as a fundamental tool, and the problem of finding its structure from data has been received much attention in many fields including social sciences [1], bioinformatics [9] and neuroinformatics [8].

Among variety of models, structural equation models (SEMs) and Bayesian networks (BNs) have been widely used to analyze causal relationships in empirical studies [1, 10, 13]. However, the full structure, *i.e.*, a causal ordering and connection strengths, of the model cannot be identified in most cases without prior knowledge on the structure when only covariance structure of data is used for model estimation as is the case in almost conventional methods. Recently, it is reported that non-Gaussian structure of data overcomes this identifiability problem in the case of linear directed acyclic graphs (DAGs) [11, 12]. By their algorithms (the LiNGAM algorithms), if the external influences are non-Gaussian, the structure can be uniquely estimated by using observed data only without any prior knowledge (under an assumption of acyclicity).

However, the applicability of the LiNGAM algorithms might be restricted in some real-world applications because of its relatively strong assumptions of linear acyclicity in each variable. For example, in the cases where an unobserved confounder exists between exogenous variables or sink variables, a DAG structure is no longer appropriate to apply. Thus, it would be useful to develop a non-Gaussianity based framework to estimate the structure of more general class of models, such as chain graphs [7], so as to deal with the situations under which the assumptions on

[1]The Institute of Scientific and Industrial Research (ISIR), Osaka University.
E-mail: {kawahara,sshimizu,washio}@ar.sanken.osaka-u.ac.jp
[2]Department of Sociology, The University of North Carolina.
E-mail: bollen@unc.edu

DAGs are not satisfied. Note that there is a method based on non-Gaussianity that takes unobserved confounders into account [4]. However, since this method needs to model unobserved variables explicitly, the computational cost is crucially high (In fact, only two or three variables were empirically treated in the paper).

In this paper, we propose a non-Gaussian variant of chain graphs, which includes the one of linear acyclic graphs as a special case, and present an algorithm for the estimation of this model. The algorithm finds an ordering of the subsets of variables by iteratively evaluating the independence between the variable subset and the residuals when the remaining variables are regressed on those. In addition to the applicability to chain graphs, it is empirically verified that the estimation by the proposed algorithm works reasonably well compared with the existing algorithms when applied to DAGs. However, this procedure needs to compute the independence exponentially many times corresponding to the number of variables. Therefore, we propose an approximate approach that can be performed without depending on the number of variables (although the accuracy may depend on) and can be applied to large scale graphs. The performance will be illustrated using artificial and real-world datasets.

The reminder of this paper is organized as follows. In Sect. 2, we first introduce a linear non-Gaussian acyclic model for sets of variables (GroupLiNGAM model). Then in Sect. 3, we present an algorithm for (directly) estimating the GroupLiNGAM model. However, this approach would be inefficient for large sized graphs. Therefore, in Sect. 4, we give an approximate approach that can be applied to large sized graphs based on the algorithm described in Sect. 3. The algorithms are illustrated and examined in its performance using artificial data in Sect. 5 and real-world data in Sect. 6. Finally, we give conclusions in Sect. 7.

2. GroupLiNGAM model

In this paper, we consider a non-Gaussian variant of chain graphs, which we call the *GroupLiNGAM model*.

Assume that observed data are generated from a process represented graphically by a chain graph on random variables $\boldsymbol{x}$ of dimension p. Let us express this chain graph by a $p \times p$ adjacency matrix $B = \{b_{ij}\}$, where every b_{ij} represents the connection strength from a variable x_j to another x_i in the chain graph. Also, let $K(l)$ $(l = 1, \dots, m,\ m \leq p)$ be ordered blocks, *i.e.*, disjoint subsets of variables, so that no variables in later subsets influence any variable in earlier subsets and $K(1) \cup \cdots \cup K(m) = V$, where $V := \{1, \dots, p\}$ is the indices set of the variables.[3] The index of the subset, *i.e.*, l, that x_i belongs to will be referred as $l(i)$. Moreover, assume that the relations between variables in different subsets are linear. Without loss of generality, each observed variable x_i is assumed to have zero mean.

Then, the GroupLiNGAM model is represented as

$$x_i = \sum_{l(j) \leq l(i), i \neq j} b_{ij} x_j + e_i, \tag{1}$$

where e_i is an external influence. All external influences e_i's are non-Gaussian random variables with zero means and non-zeros variances, and independent of

[3] This definition is a generalization of the one of a DAG. That is, this is actually the definition of a DAG if all the subsets consist of one element, *i.e.*, $m = p$.

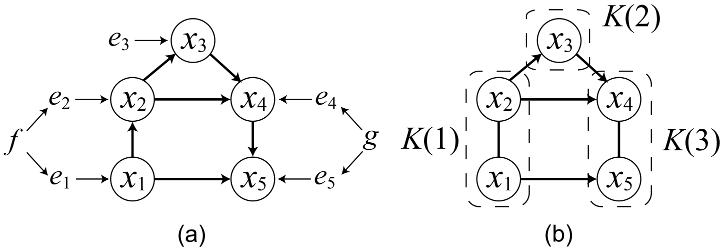

FIGURE 1. Illustrative example of a chain graph.

each other in different blocks. Alternatively, we write the model (1) in a matrix form:

$$\boldsymbol{x} = B\boldsymbol{x} + \boldsymbol{e}, \tag{2}$$

where B can be permuted by simultaneous equal row and column permutations to be lower block-triangular due to the acyclicity of disjoint subsets in chain graphs [14, 2]. Moreover, if we represent the model (2) as

$$\boldsymbol{x} = A\boldsymbol{e}, \tag{3}$$

the matrix A $(:= (I - B)^{-1})$ (called a mixing matrix) also becomes lower block-triangular (and with all unities in the diagonal). Note that, in the case of $m = p$, *i.e.*, the DAG case, the model (3) defines the independent component analysis (ICA) model [5] since the components of $\boldsymbol{e}$ are independent and non-Gaussian. Since the ICA model is identifiable, the model (3) in this case $(m = p)$ is also identifiable, which is the key idea of the original LiNGAM algorithm [11] (we call it ICA-LiNGAM in the later part of this paper).

Now, let us consider an illustrative example in which the model is represented by (cf. Figure 1 (a))

$$\begin{aligned} x_1 &= e_1, \\ x_2 &= b_{21}x_1 + e_2, \\ x_3 &= b_{32}x_2 + e_3, \\ x_4 &= b_{42}x_2 + b_{43}x_3 + e_4, \\ x_5 &= b_{51}x_1 + b_{54}x_4 + e_5, \end{aligned} \tag{4}$$

where unobserved confounders f and g exist between e_1 and e_2 and between e_4 and e_5, respectively, as

$$e_1 = c_1 f + d_1 \text{ and } e_2 = c_2 f + d_2,$$

and

$$e_4 = c_4 g + d_4 \text{ and } e_5 = c_5 g + d_5.$$

d_1, d_2, d_4 and d_5 are independent of each other. Note that, in this case, the assumption for the LiNGAM algorithms, *i.e.*, exogenous influences are independent of each other, is not satisfied. In fact, x_1 and x_4 depend on x_2 and x_5, respectively, because f and g are not observed, and a DAG representation is no longer appropriate to apply. The ordered blocks for the example (4) are $K(1) = \{1, 2\}$, $K(2) = \{3\}$ and $K(3) = \{4, 5\}$ (cf. Figure 1 (b)).

3. Model estimation

In this section, we address the estimation of the GroupLiNGAM model from data. In the following parts, we will refer the subset of variables corresponding to $S \subseteq V$ as $\boldsymbol{x}_S$. Also, we denote by $V \setminus S$ the complementary set of V with respect to S, and by $\boldsymbol{x}_{\bar{S}}$ the subset of variables corresponding to $V \setminus S$.

3.1. Identifying exogenous variables using non-Gaussianity.

Recently, it has been reported that non-Gaussianity of external influences serves for directly estimating the ordering of variables from data [12] (DirectLiNGAM). The key insight herein is that, once an exogenous variable is identified, we can remove the component of the exogenous variable from the other variables without violating the original ordering for the residuals when the exogenous variable is regressed on the remaining variables. Here, we describe the analogous insight still holds for sets of variables. To this end, we first need the following assumption:

Definition 1 (correlation-faithfulness)**.** *The distribution of $\boldsymbol{x}$ is said to be correlation-faithful to the generating graph if correlation and conditional correlation of x_i are entailed by the graph structure, i.e., the zeros/non-zeros status of b_{ij}, but not by specific parameter values of b_{ij}.*

This concept is motivated by the faithfulness [13]. Also, we give the definition of the exogenous set of variables as follows.

Definition 2 (exogenous set)**.** *Let the partition of the variables $\boldsymbol{x}$ be $\boldsymbol{x} = (\boldsymbol{x}_S, \boldsymbol{x}_{\bar{S}})$ such that $\boldsymbol{x}_S$ and $\boldsymbol{x}_{\bar{S}}$ are not empty. Then, the subset of variables $\boldsymbol{x}_S$ is said to be exogenous against $\boldsymbol{x}_{\bar{S}}$, if the corresponding partition of the the matrix B has the following form:*

$$B = \begin{bmatrix} B_S & 0 \\ B_{\bar{S},S} & B_{\bar{S}} \end{bmatrix}.$$

Note that each variable in the exogenous set is not necessarily an exogenous variable. That is, the variables in an exogenous set may be influenced by each other inside of the set. Also note that the submatrix of the mixing matrix A corresponding to B_S is full-rank because the covariance matrix Σ_S of $\boldsymbol{x}_S$ is also full-rank from the correlation-faithfulness assumption.

Now, we give two lemmas and one corollary that is the basis of the algorithm proposed in this paper.

Lemma 3. *Assume that the input data $\mathbf{x}$ follows the GroupLiNGAM model* (2)*, and that the distribution of $\boldsymbol{x}$ is correlation-faithful to the generating graph. Let $r^{(S)}$ be the residual when $x_{\bar{S}}$ is regressed on $\boldsymbol{x}_S$ for $S \subset V$: $r^{(S)} = x_{\bar{S}} - \Sigma_{S,\bar{S}}^T \Sigma_S^{-1} \boldsymbol{x}_S$, where*

$$\Sigma = \begin{bmatrix} \Sigma_S & \Sigma_{S,\bar{S}} \\ \Sigma_{S,\bar{S}}^T & \Sigma_{\bar{S}} \end{bmatrix}$$

is the covariance matrix of $(\boldsymbol{x}_S, \boldsymbol{x}_{\bar{S}})$. Then, a set of variables $\boldsymbol{x}_S$ is exogenous if and only if $\boldsymbol{x}_S$ is independent of its residual $\boldsymbol{r}^{(S)}$.

Proof First, assume that $\boldsymbol{x}_S$ is exogenous. Then, one can write $\boldsymbol{x}_{\bar{S}} = A_{\bar{S},S} A_S^{-1} \boldsymbol{x}_S + \bar{\boldsymbol{e}}_{\bar{S}}^{(S)}$, where

$$A = \begin{bmatrix} A_S & 0 \\ A_{\bar{S},S} & A_{\bar{S}} \end{bmatrix}$$

is the coefficient matrix in Eq. (3). From the definition of model (3), $\bar{\boldsymbol{e}}^{(S)}_{\bar{S}} = A_{\bar{S}}\boldsymbol{e}_{\bar{S}}$ and $\boldsymbol{x}_S$ are mutually independent. Also, since $\Sigma^T_{S,\bar{S}} = A_{\bar{S},S}A_S^{-1}\Sigma_S$, $A_{\bar{S},S}A_S^{-1}$ is equivalent to the regression coefficients when $\boldsymbol{x}_{\bar{S}}$ is regressed on $\boldsymbol{x}_S$. Therefore, $\boldsymbol{r}^{(S)}$ is equivalent to $\bar{\boldsymbol{e}}^{(S)}_{\bar{S}}$. As a result, $\boldsymbol{x}_S$ and $\boldsymbol{r}^{(S)}$ are mutually independent. Next, assume that $\boldsymbol{x}_S$ is independent of $\boldsymbol{r}^{(S)}$. Then, since $\boldsymbol{x}_S$ is independent of $\boldsymbol{e}_{\bar{S}}$, all elements of the regression coefficient matrix when $\boldsymbol{x}_S$ is regressed on $\boldsymbol{e}_{\bar{S}}$, *i.e.*, $A_{S,\bar{S}}$, become zeros, which means all elements of the upper right part of B, *i.e.*, $B_{S,\bar{S}}$, are also zeros. From the correlation-faithfulness assumption and the definition of exogeneous sets, $\boldsymbol{x}_S$ is exogenous. ■

Lemma 4. *Assume the assumptions of Lemma 3 and that a set of variables* $\boldsymbol{x}_S$ *is exogenous. Let* $\boldsymbol{r}^{(S)}$ *be the residual vector when* $\boldsymbol{x}_{\bar{S}}$ *is regressed on* $\boldsymbol{x}_S$ *for* $S \subset V$*. Then, GroupLiNGAM models hold both for* $\boldsymbol{x}_S$ *and* $\boldsymbol{r}^{(S)}$*, respectively :* $\boldsymbol{x}_S = B_S\boldsymbol{x}_S + \boldsymbol{e}_S$ *and* $\boldsymbol{r}^{(S)} = B^{(S)}\boldsymbol{r}^{(S)} + \mathbf{e}^{(S)}$*, where* B_S *and* $B^{(S)}$ *are matrices that can be permuted to be block lower-triangular by simultaneous row and column permutations, and elements of* $\boldsymbol{e}_S$ *and* $\mathbf{e}^{(S)}$ *are non-Gaussian and mutually independent in different blocks, respectively.*

Proof Without loss of generality, assume that B in the GroupLiNGAM model (2) is already permuted to be lower block-triangular (which means A is also to be lower block-triangular with all unities in the diagonal). First, it is straightforward from Def. 2 that B_S is lower block-triangular. Next, since $\boldsymbol{x}_S$ is exogenous, the regression coefficients when $\boldsymbol{x}_{\bar{S}}$ is regressed on $\boldsymbol{x}_S$ becomes $A_{\bar{S},S}A_S^{-1}$. Therefore, removing the effects of $\boldsymbol{x}_S$ from $\boldsymbol{x}_{\bar{S}}$ by least-squares estimation is equivalent to setting all elements of the first $|S|$ columns of A to be zeros. This means that the residuals $\boldsymbol{r}^{(S)}$ are not influenced by $\boldsymbol{x}_S$ because of the correlation-faithfulness assumption. As a result, we again obtain a lower block-triangular mixing matrix with all unities in the diagonal $A^{(S)}(= A_{\bar{S}})$ for $\boldsymbol{r}^{(S)}$. ■

Corollary 5. *Assume the assumptions in Lemma 4. Denote by* $l_S(i)$ *and* $l_{\boldsymbol{r}^{(S)}}(i)$ *the indices of the ordered subsets encoded by the chain graphs on* $\boldsymbol{x}_S$ *and* $\boldsymbol{r}^{(S)}$*, respectively. Recall that* $l(i)$ *denotes the index of the ordered subsets encoded by the chain graph on* $\boldsymbol{x}$*. Then, the ordering of the subsets of* $\boldsymbol{x}_S$ *and* $\boldsymbol{r}^{(S)}$ *are respectively equivalent to that of corresponding original subsets of variables, i.e.,* $l_S(i_1) < l_S(i_2) \Leftrightarrow l(i_1) < l(i_2)$ *and* $l_{\boldsymbol{r}^{(S)}}(i_1) < l_{\boldsymbol{r}^{(S)}}(i_2) \Leftrightarrow l(i_1) < l(i_2)$*.*

Proof As described in the proof of Lemma 4, the adjacency matrices (and the mixing matrices) for the GroupLiNGAM models on $\boldsymbol{x}_S$ and $\boldsymbol{r}^{(S)}$ are equivalent to the corresponding parts of the one for the GroupLiNGAM model on $\boldsymbol{x}$. This shows the orderings of $\boldsymbol{x}_S$ and $\boldsymbol{r}^{(S)}$ are not changed. ■

Lemma 3 indicates that an exogenous set is identified by evaluating the independence between a set of variables $\boldsymbol{x}_S$ and its residuals $\boldsymbol{r}^{(S)}$. Lemma 4 implies that the GroupLiNGAM models for the p-dimensional vector $\boldsymbol{x}_S$ and the $(p - |S|)$-dimensional residual vector $\boldsymbol{r}^{(S)}$ can be handled as new input models, and Lemma 3 can be further applied to the each model to derive the next set of exogenous variables. This process can be repeated until all subsets of variables are not able to be

Algorithm 1 GroupLiNGAM

```
1: Given a $p$-dimensional variables $\boldsymbol{x}$, a set of its subscripts $V$, a $p \times n$ data matrix
   of the variables $\mathbf{X}$, initialize an ordered subset of variables as $K \leftarrow \emptyset$.
2: Call $K \leftarrow$ GroupSearch ($V$, $K$, $\mathbf{X}$).
3: Construct a lower block-triangular matrix $B$ by following the order in $K$, and
   estimate the connection strengths $b_{ij}$ (using some conventional covariance-based
   regression, such as least-squares and maximum-likelihood approaches) on the
   original variables $\boldsymbol{x}$ and data matrix $\boldsymbol{X}$.

function $K \leftarrow$ GroupSearch ($U$, $K$, $\mathbf{X}_U$)
 1: for $S \subset U$ do
 2:     Perform least-squares regression of $\boldsymbol{x}_S$ on $\boldsymbol{x}_{U \setminus S}$ (denote the residual vector
        by $\boldsymbol{r}^{(S)}$ and its residual data matrix by $\mathbf{R}^{(S)}$) and then compute some inde-
        pendence measure $I(S)$ between $\boldsymbol{x}_S$ and $\boldsymbol{r}^{(S)}$, e.g., $MI(\boldsymbol{x}_S, \boldsymbol{r}^{(S)})$.
 3: end for
 4: $S_* := \arg\min I(S)$.
 5: if $I(S_*) \leq \delta$ and $|U| \neq 1$ then
 6:     Set $\mathbf{X}_{U \setminus S_*} \leftarrow \mathbf{R}^{(S_*)}$.
 7:     Call $K \leftarrow$ GroupSearch $(S_*, K, \mathbf{X}_{S_*})$
 8:     Call $K \leftarrow$ GroupSearch $(U \setminus S_*, K, \mathbf{X}_{U \setminus S_*})$
 9: else
10:     Append $S_*$ to the end of $K$.
11: end if
```

devided, and the resulting order of the sets of variable subscripts shows the causal order of the original observed variables according to Corollary 5.

As the independence measure used in Lemma 3, the mutual information between the subset of variable and the residuals, *i.e.*, $MI(\boldsymbol{x}_S, \boldsymbol{r}^{(S)})$, would be available. There are many options for its estimation from data. In the later experiments, we used an algorithm based on the k-nearest neighbors method [6].[4] This method has one tuning parameters, *i.e.*, the number of neighbors *kneig*. Although the setup of this parameter is not trivial, the algorithm is known to work well empirically when *kneig* is set as 3–5 % of the dimensional p [6].

3.2. **GroupLiNGAM algorithm.** Based on the above result, we now present an algorithm to estimate a block causal ordering and the connection strengths in the GroupLiNGAM model under the correlation-faithfulness assumption. The pseudocode of the algorithm is shown in Alg. 1.

The algorithm is performed by the recursive calls of GroupSearch function, which devides a given subset U into ordered two groups. Since an exogenous set is identified by evaluating the independence between a subset of U and its residuals from Lemma 3, we find such subset $S_*(\subset U)$ as the one that minimizes some independence measure $I(S)$ (Lines 1–3 in GroupSearch in Alg. 1). Thus, U is divided into ordered two groups S_* and $U \setminus S_*$. From Lemma 4, for each of S_* and $U \setminus S_*$, the GroupLiNGAM models hold. Therefore, this procedure is iterated until further partition cannot be found, which is judged with a threshold δ (Line 8–9 in

[4] We used the MATLAB code available from `http://www.klab.caltech.edu/~kraskov/MILCA/` in the experiments.

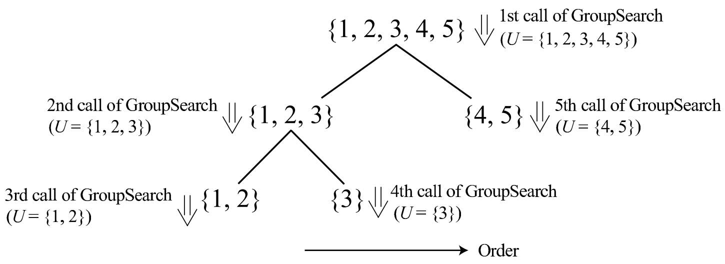
{1, 2, 3, 4, 5}
1st call of GroupSearch (U = {1, 2, 3, 4, 5})
2nd call of GroupSearch (U = {1, 2, 3})
{1, 2, 3}
{4, 5}
5th call of GroupSearch (U = {4, 5})
3rd call of GroupSearch (U = {1, 2})
{1, 2}
{3}
4th call of GroupSearch (U = {3})
Order

Figure 2. Illustration of Alg. 1 for the example (4).

GroupSearch in Alg. 1). Finally-obtained order of variable subsets is consistent globally, which is guaranteed by Corollary 5. The illustration of this procedure for the example Eq. (4) is shown in Fig. 2.

Note that Alg. 1 is specialized to the DAG case if we set $\delta = +\infty$. However, the outputs by Alg. 1 and the DirectLiNGAM algorithm are not always same because Alg. 1 finds the subset of variables that is exogenous against the remaining variables while the DirectLiNGAM algorithm identifies an exogenous variable iteratively (that is, we can say that the former uses global information of independence between variables while the latter local one). Thus, the accumulation of errors of regression in Alg. 1 is expected to be no more than the one in the DirectLiNGAM algorithm, which will be illustrated empirically in Sect. 5.

4. Approximate approach for large graphs

Since Alg. 1 needs to compute independence between $\boldsymbol{x}_S$ and $\boldsymbol{r}^{(S)}$ exponentially many times ($2^{|U|-1}$, once for every $S \subset U$)[5] at each iteration (Lines 1–3 in GroupSearch), it can be applied only to medium sized graphs (consisting of up to around 15 nodes). Here, we propose an approximate approach based on Alg. 1 applicable to larger sized graphs (L-GroupLiNGAM).

The basic idea of the proposed algorithm is as follows. If we observe only a subset of variables, then some of the unobserved variables may act as confounders against some of the observed variables and, as a result, causal directions between such observed variables have become not identifiable. However, since the definition of our model permits confounders, *i.e.*, makes edges undirected if there exist confounders between the observed variables, we can find the order of blocks which is identifiable from the currently observed variables using Alg. 1. Therefore, by randomly picking up the subset of variables such that the set of the subsets covers all variables and applying Alg. 1 to each subset, we finally obtain a block ordering of all variables in a large graph. The validness of this procedure can be guaranteed by the following proposition:

Proposition 6. *Assume the assumptions in Lemma 4. Let denote by $\tilde{l}_T(i)$ $(T \subset V)$ the ordering of the subsets of variables when only $\boldsymbol{x}_T$ is observed (and other variables $\boldsymbol{x}_{\bar{T}}$ are not observed). Then, the order $\tilde{l}_T(i)$ is consistent with the one when all variables are observed, i.e., $\tilde{l}_T(i_1) < \tilde{l}_T(i_2) \Rightarrow l(i_1) < l(i_2)$.*

Proof Assume that only $\boldsymbol{x}_T$ is observed for $T \subset V$. If $\tilde{l}_T(i_1) < \tilde{l}_T(i_2)$, then there exists a subset $S \subset T$ exogenous against $T \setminus S$ such that $i_1 \in S$ and $i_2 \in T \setminus S$.

[5] $U = V$ $(|V| = p)$ at the first iteration.

Algorithm 2 L-GroupLiNGAM

```
1: Given a $p$-dimensional variables $\boldsymbol{x}$, a set of its subscripts $V$, a $p \times n$ data matrix
   of the variables $\mathbf{X}$ and a cardinality $h$, initialize the list of orders between
   combinations of variables $\tilde{k} = \emptyset$.
2: Compute a random covering $T(i)$ $(i = 1, \ldots, N)$ of variables with cardinarity
   $h$.
3: for $i = 1, \ldots, N$ do
4:     Apply Alg. 1 modefied by replacing Line 1 in GroupSearch func. to (5) with
       $V \leftarrow T(i)$, and add new orders from its output $K$ to $\tilde{k}$.
5: end for
6: Construct a block order $\tilde{K}$ for all variables from $\tilde{k}$.
7: Construct a strictly lower block-triangular matrix $B$ by following the order in $\tilde{K}$,
   and estimate the connection strengths $b_{ij}$ (using some conventional covariance-
   based regression, such as least-squares and maximum-likelihood approaches) on
   the original variables $\boldsymbol{x}$ and data matrix $\boldsymbol{X}$.
```

Therefore, one can write $\boldsymbol{x}_S = \tilde{A}_S \tilde{\boldsymbol{e}}_S$ and $\boldsymbol{x}_{T \setminus S} = \tilde{A}_{T \setminus S, S} \tilde{A}_S^{-1} \boldsymbol{x}_S + \tilde{A}_{T \setminus S} \tilde{\boldsymbol{e}}_{T \setminus S}$, where $\boldsymbol{x}_S$ and $\tilde{A}_{T \setminus S} \tilde{\boldsymbol{e}}_{T \setminus S}$ are mutually independent. This means, if we denote as $\tilde{\boldsymbol{e}}_S = \sum_{i \in S_1} \boldsymbol{a}_{S,i} e_i$ and $\tilde{\boldsymbol{e}}_{T \setminus S} = \sum_{i \in S_2} \boldsymbol{a}_{T \setminus S, i} e_i$, where $S_1 \subseteq S \cup (V \setminus T)$ and $S_2 \subseteq (T \setminus S) \cup (V \setminus T)$, then the union of S_1 and S_2 is empty, *i.e.*, $S_1 \cap S_2 = \emptyset$. This means that all elements of the submatrix $\{a_{ij}\}$ $(i \in S_1 \cup S, j \in V \setminus (S_1 \cup S))$ of the mixing matrix A are zeros and, as a result, $l(i_1) < l(i_2)$. ■

Based on the above result, we now present an algorithm for estimating the GroupLiNGAM model with large number of variables. The pseudo-code of the algorithm is shown in Alg. 2, where $\tilde{k}$ is the list of combinations of variables (j_1, j_2) with orders $l(j_1) < l(j_2)$.

In the algorithm, we first generate a random covering of all variables $T(i)$ $(i = 1, \ldots, N)$ (Line 2 in Alg. 2), *i.e.*, subsets $T(i) \subset V$ such that $\cup_{i=1,\ldots,N} T(i) = V$. And, we apply Alg. 1 to each $T(i)$ (Lines 3–5 in Alg. 2). Then, in order to reflect already-known orders (j_1, j_2) $(j_1, j_2 \in T(i))$ when choosing $S \subset U$ in Lines 1–3 in GroupSearch function in Alg. 1, we replace Lines 1 in GroupSearch to the following:

$$\textbf{for } S \subset U \text{ s.t. } j_2 \in S \to j_1 \notin U \setminus S \text{ for } (j_1, j_2) \in \tilde{k} \textbf{ do} \tag{5}$$

Also, the application of Alg. 1 may generate an output making a cycle as a whole when combined with previously-obtained orders due to statistical uncertainty of samples as the iterations of Lines 3–5 in Alg. 2 continue. In such a case, the inconsistent (old and new) orders need to be removed, *i.e.*, we merge the ordered variables by these orders into a group. Finally, the ordering of variable subsets is constructed from the list of obtained orders. Although this procedure may not be able to find some of block orders, more and more ones are expected to be found depending on subsets $T(i)$ as the iteration continue.

5. Simulations

In this section, we evaluate the proposed algorithms empirically using artificial datasets. Especially, we focus on (i) the evaluation of the validity of the proposed algorithms for estimating the GroupLiNGAM model (Alg. 1 and Alg. 2) and (ii)

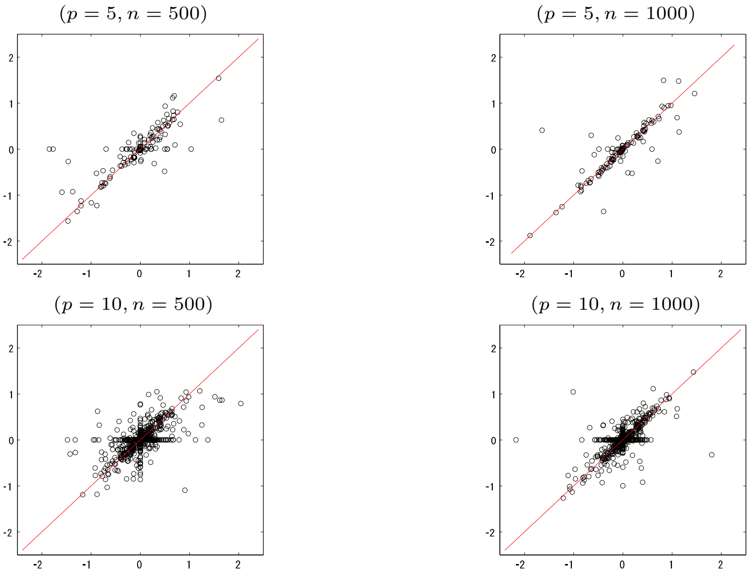

FIGURE 3. Scatter-plots of the estimated b_{ij} by Alg. 1 (vertical axis) versus the generating values (horizontal axis) for combinations of dimensionality $p = (5, 10)$ and the number of samples $n = (500, 1000)$.

the comparison of estimation accuracy of the proposed and existing algorithms (ICA-LiNGAM [11] and DirectLiNGAM [12]) in DAG cases.

First, for the purpose (i), we created datasets under each combination of number of variables p, sample size n and coverage cardinality h (for Alg. 2), as follows.[6]

(1) First, a $p \times p$ block lower-triangular matrix B was randomly created so that the standard deviations of variables owing to their parents (determined from its ordering of the subsets of variables) ranged in the interval $[0.5, 1.5]$, where the number of blocks and the maximum number of parents of the created network for B were also randomly determined from 1 to p in uniform manner. The standard deviations of the external influences $\boldsymbol{e}$ were randomly selected from the interval $[0.5, 1.5]$.

(2) Next, we generated data with sample size n by independently drawing the external influence variables $\boldsymbol{e}$ from various non-Gaussian distributions with zero mean and unit variance. This is performed by generating Gaussian variables z_i with zero means and unit variances, transformed it as $e_i = \text{sign}(z_i)|z_i|^{q_i}$, where nonlinear exponents q_i were randomly selected from the interval $[0.5, 0.8] \cup [1.2, 2.0]$,[7] and then standardizing e_i to have zero means and unit variables.

(3) The values of the observed variables $\boldsymbol{x}$ were generated according to the GroupLiNGAM model (2). And, the order of $\boldsymbol{x}$ is permuted randomly.

[6] The way of creating datasets is the same as [12] except for that B is a block lower-triangular.

[7] Nonlinear exponents q_i with $[0.5, 0.8]$ and $[1.2, 2.0]$ give sub-Gaussian and super-Gaussian variables, respectively.

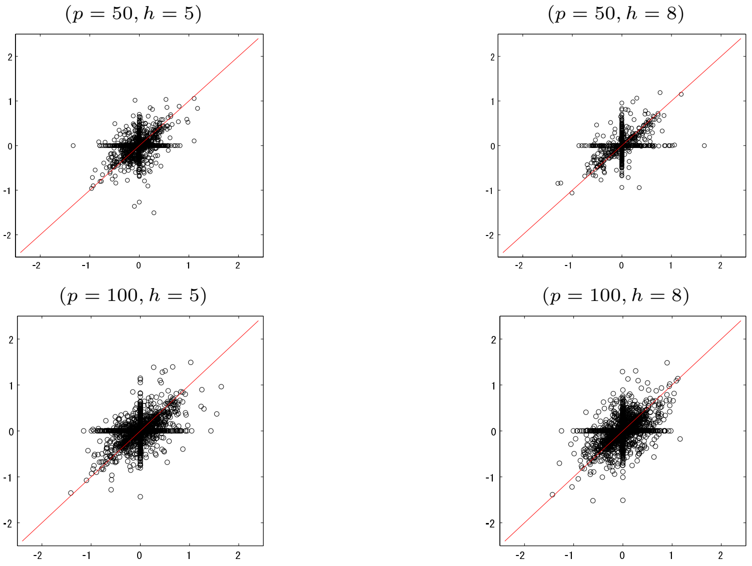

Figure 4. Scatter-plots of the estimated b_{ij} by Alg. 2 (vertical axis) versus the generating values (horizontal axis) for combinations of dimensionality $p = (50,\ 100)$ and coverage cardinality $h = (5, 8)$ $(n = 1000)$.

The graphs in Fig. 3 and Fig. 4 show the scatter-plots of the elements of the estimated and generating adjacency matrix B (for randomly generated 10 datasets in the respective case). GroupLiNGAM (Alg. 1) and L-GroupLiNGAM (Alg. 2) were respectively applied to relatively small and large sized graphs (p=5, 10 (Fig. 3) and p=50, 100 (Fig. 4)). The parameters δ and *kneig* were set as 1.0×10^{-2} and $0.05 \times n$, respectively. For Alg. 2, the number of subsets in a covering, *i.e.*, N, was set as 50 in the experiment. Although the estimation seems to fail sometimes depending on dimensionality p, the number of samples n or coverage cardinality h, the estimation seems to work reasonably well.

Next, for the purpose (ii), we created datasets as in the same manner with the procedure described in [11], which is same with the above procedure except that each block contains only one element. As described above, GroupLiNGAM is specialized to the DAG case by setting δ as $+\infty$ and thus, in this experiment, δ was set as 1×10^6 for Alg. 1. The graphs in Fig. 5 show the medians of the numbers of errors, *i.e.*, the numbers of elements in the strictly upper triangular part when the *true* connection strength matrix B is permuted according to the estimated orders by the algorithms. The median errors by the algorithms are similar for almost experimental conditions and, hence, we can say that the estimation of GroupLiNGAM works reasonably well in the DAG case too. Here, we should note again that GroupLiNGAM can be applied not only to DAGs but also to chain graphs while the existing algorithms (ICA-LiNGAM and DirectLiNGAM) cannot.

6. Application to real data

To evaluate the applicability of the proposed algorithm (GroupLiNGAM), we analyzed a dataset taken from a sociological data repository on the Internet called

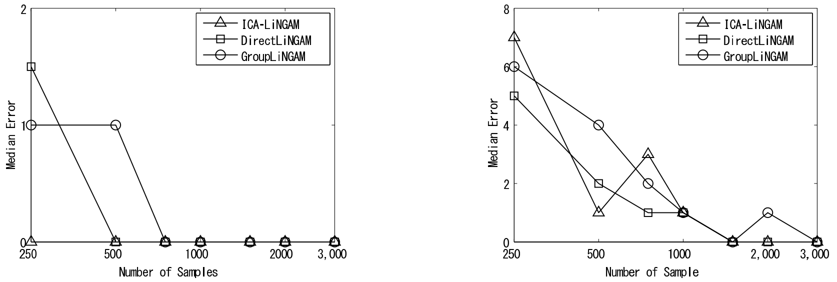

FIGURE 5. Median numbers of errors in estimated orders by the existing and proposed algorithms when applied to DAG cases (Left: $p = 6$ and Right: $p = 10$).

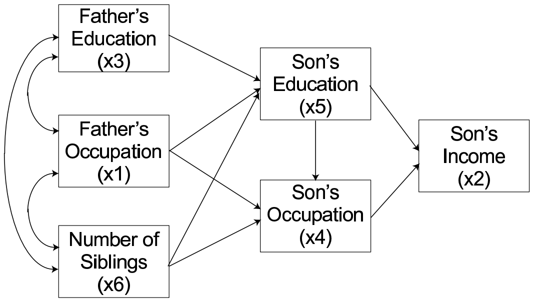

FIGURE 6. Status attainment model based on domain knowledge, where $\{1,3,6\}<\{5\}<\{4\}<\{2\}$.

General Social Survey.[8] The data consisted of six observed variables, x_1: father's occupation level, x_2: son's income, x_3: father's education, x_4: son's occupation level, x_5: son's education, x_6: number of siblings. The sample size was 1,380. Fig. 6 shows domain knowledge about their causal relations: $K(1)=\{1,3,6\}$, $K(2)=\{5\}$, $K(3)=\{4\}$ and $K(4)=\{2\}$. In this section, we represent such relations by $\{1,3,6\}<\{5\}<\{4\}<\{2\}$ to save space. Note that if $\{i,j\}<\{k\}$, x_i and x_j could directly and/or indirectly cause x_k, but not vice versa.

In this experiment, Alg. 1 was applied since the number of variables is small. We tested several numbers of nearest neighbors $kneig = 40, 50, 60, 70$ to compute mutual information using the k-nearest neighbor approach [6] for GroupLiNGAM. The estimated networks were not sensitive to the choice of the number of nearest neighbors $kneig$, and essentially the same results were obtained for the values of $kneig$. We show the results under $kneig$=50 in Tab. 1, where a smaller threshold value for independence δ gives a network between larger groups of variables.

We first analyzed all the six variables. The estimated orders by ICA-LiNGAM [11], Direct-LiNGAM [12] and GroupLiNGAM are shown at the second top of Tab. 1. Those estimated orders are difficult to interpret since son's income (x_2) and/or son's education (x_5) could cause father's variables (x_1,x_3), but not vice versa. The orders are not reasonable to their temporal orderings.

[8] http://www.norc.org/GSS+Website/

Domain knowledge:	$\{1,3,6\}<\{5\}<\{4\}<\{2\}$
All the six variables analyzed.	
ICA-LiNGAM:	$\{5\}<\{6\}<\{3\}<\{1\}<\{4\}<\{2\}$
DirectLiNGAM:	$\{6\}<\{2\}<\{1\}<\{3\}<\{4\}<\{5\}$
GroupLiNGAM:	
δ=0.500	$\{6\}<\{2\}<\{1\}<\{4\}<\{5\}<\{3\}$
δ=0.100	$\{6\}<\{2\}<\{1\}<\{4,5\}<\{3\}$
δ=0.010	$\{6\}<\{2\}<\{1,3,4,5\}$
δ=0.001	$\{1,2,3,4,5,6\}$
x_2 omitted.	
ICA-LiNGAM:	$\{5\}<\{6\}<\{3\}<\{1\}<\{4\}$
DirectLiNGAM:	$\{6\}<\{1\}<\{3\}<\{4\}<\{5\}$
GroupLiNGAM:	
δ=0.500	$\{6\}<\{1\}<\{3\}<\{5\}<\{4\}$
δ=0.100	$\{6\}<\{1,3\}<\{5\}<\{4\}$
δ=0.050	$\{6\}<\{1,3\}<\{4,5\}$
δ=0.010	$\{6\}<\{1,3,4,5\}$
δ=0.001	$\{1,3,4,5,6\}$
x_2 and x_6 omitted.	
ICA-LiNGAM:	$\{5\}<\{3\}<\{1\}<\{4\}$
DirectLiNGAM:	$\{1\}<\{3\}<\{4\}<\{5\}$
GroupLiNGAM:	
δ=0.50	$\{3\}<\{1\}<\{5\}<\{4\}$
δ=0.10	$\{1,3\}<\{5\}<\{4\}$
δ=0.05	$\{1,3\}<\{4,5\}$
δ=0.01	$\{1,3,4,5\}$

Table 1. Estimated orders of groups.

Next, we omitted son's income (x_2) and analyzed the other five variables. Omitting x_2 would not create any unobserved confounder since it does not cause any other variables according to the domain knowledge. The results are shown in the third top of Tab. 1. DirectLiNGAM and GroupLiNGAM found consistent time orderings between father's variables (x_1,x_3) and son's variables (x_4,x_5). Furthermore, GroupLiNGAM found a reasonable ordering between son's variables, *i.e.*, son's education (x_5) could cause son's occupation level (x_4), but not vice versa, whereas DirectLiNGAM failed. However, number of siblings (x_6) is the top variable in every estimated ordering by DirectLiNGAM and GroupLiNGAM and could cause father's variables (x_1,x_3), which is not easy to interpret.

We further omitted number of siblings (x_6) as well as son's income (x_2) and analyzed the other four variables (x_1,x_3,x_4,x_5). Omitting x_6 could create an unobserved confounder since it could relate father's variables (x_1,x_3) and son's variables (x_4,x_5). The bottom of Tab. 1 shows the results. Every ordering estimated by GroupLiNGAM is consistent with the domain knowledge. ICA-LiNGAM wrongly estimated that son's education (x_5) could cause father's variables (x_1,x_3), but not vice versa. DirectLiNGAM also gave inconsistent orderings between father's education (x_3) and father's occupation (x_1) and between son's education (x_5) and son's occupation (x_4).

In summary, GroupLiNGAM provided more consistent orderings with the domain knowledge than ICA-LiNGAM and DirectLiNGAM. The reason would be that only GroupLiNGAM is able to allow unobserved confounders. However, it is not yet very clear why the inclusion of x_2 and x_6 makes the results difficult to interpret. One possibility is that x_2 and x_6 might not fit well some assumption in the three discovery methods, *e.g.*, linearity, compared to the other four variables.

7. Conclusions

In this paper, we proposed the GroupLiNGAM model, a non-Gaussian variant of chain graphs, and presented an algorithm for estimating this model, which is identifiable without any prior knowledge on the structure. Based on the result that an exogeneous set is identified by evaluating the independence between a variable subset and the residuals when the remaining variables are regressed on those, the proposed algorithm finds an ordered devision of variables iteratively and identifies an ordering of disjoint subsets of variables. However, since the computational cost grows exponentially according to the number of variables, a middle sized graph is the practical limit of this algorithm. Therefore, in addition, we presented an approximate approach to apply this framework to large sized graphs. In the experimental parts, we evaluated the algorithms empirically and illustrated the applicability using artificial and real datasets.

The algorithm has a tuning parameter δ, which determines when the devision of groups should be stopped (*kneig* is also an tuning parameter in the current implementation. However, this parameter is for the estimation of mutual information by the k-nearest neighbor method [6] and thus would not be an essential parameter in our method). For more exact devision of groups, it would be useful to combine our framework with some statistical test method, such as the bootstrap method [3], in the future. Also, in the current implementation, exponentially large number of subsets need to be examined when identifying an exogenous (Line1–3 in GroupSearch in Alg. 1). Therefore, it would be important to develop more efficient search strategy for this part using some discrete structure.

References

[1] K. A. Bollen, *Structural equations with latent variables*, John Wiley & Sons, 1989.

[2] D. R. Cox and N. Wermuth, *Multivariate dependencies: Models, analysis and interpretation*, Chapman & Hall/CRC, 1996.

[3] B. Efron and R. J. Tibshirani, *An introduction to the bootstrap*, Chapman & Hall, 1994.

[4] P. O. Hoyer, S. Shimizu, A. Kerminen, and M. Palvianen, *Estimation of causal effects using linear non-gaussian causal models with hidden variables*, International Journal of Approximate Reasoning **49** (2008), no. 2, 362–378.

[5] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent component analysis*, Wiley, New York, 2001.

[6] A. Kraskov, H. Stögbauer, and P. Grassberger, *Estimating mutual information*, Physical Review E **69** (2004), no. 6, 066138.

[7] S. L. Lauritzen, *Graphical models*, Oxford University Press, 1996.

[8] A. Londei, A. D'Ausilio, D. Basso, and M. O. Belardinelli, *A new method for detecting causality in fMRI data of cognitive processing*, Cognitive Processing **7** (2006), no. 1, 42–52.

[9] R. Opgen-Rhein and K. Strimmer, *From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data*, BMC Systems Biology **1** (2007), 1–37.

[10] J. Pearl, *Causality: Models, reasoning, and inference*, Cambridge University Press, 2000.

[11] S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. Kerminen, *A linear non-gaussian acyclic model for causal discovery*, Journal of Machine Learning Research **7** (2006), 2003–2030.

[12] S. Shimizu, A. Hyvärinen, Y. Kawahara, and T. Washio, *A direct method for estimating a causal ordering in a linear non-gaussian acyclic model*, Proc. of the 25th Ann. Conf. on Uncertainty in Artificial Intelligence (UAI09), 2009, pp. 506–513.

[13] P. Spirtes, C. Glymour, and R. Scheines, *Causation, prediction and search*, 2nd ed., MIT Press, 2001.

[14] N. Wermuth and S. L. Lauritzen, *On substantive research hypotheses, conditional independence graphs and graphical chain models*, Journal of the Royal Statistical Society. Series B (Methodological) **52** (1990), no. 1, 21–50.